Supporting Online Material for

The Transcriptome of the Sea Urchin Embryo

Manoj P. Samanta, Waraporn Tongprasit, Sorin Istrail, R. Andrew Cameron, Qiang Tu, Eric H. Davidson, Viktor Stolc*

*To whom correspondence should be addressed. E-mail: vstolc@arc.nasa.gov

Published 10 November 2006, *Science* **314**, 960 (2006)
DOI: 10.1126/science.1131898

This PDF file includes:

Other Supporting Online Material for this manuscript includes the following:
(available at www.sciencemag.org/cgi/content/full/314/5801/960/DC1)

Tables S3 to S7 as zipped archive 1131898-TableS3-7.zip

Supporting Online Material

Materials and Methods

Design of the Arrays. Tiling array probes were chosen from V0.5 release of the *S. purpuratus* draft genome sequence (Spur20050415, http://www.hgsc.bcm.tmc.edu/projects/seaurchin/*)* (*S1*). A total of 10,133,868 50-mer oligonucletide probes were selected from both strands of the entire genome including all intergenic regions. We maintained an average spacing of 10 nt between the consecutive probes (table S1). Probes with undesirable features that might have caused difficulties in hybridization were excluded by means of an algorithm described previously (*S2*). The chosen probes were dividedinto 27 arrays, each containing nearly 400,000 features. In addition to the genomic probes, the arrays included 3,000 control probes for noise estimation and data normalization purposes. Among the control probes, 2,000 were genomic and 1,000 were random sequences. The Arrayscribe program (NimbleGen Systems, Madison, WI) was applied to develop virtual masks for the arrays.

Array fabrication. The probes were synthesized on glass-based arrays by means of a maskless array synthesizer (*S3*, *S4*). Although all probes from the same scaffold were usually placed within the same array, their locations were randomized to avoid any spatial bias in the hybridization data. The arrays were hybridized with cDNA reverse transcribed from purified poly-A RNA mixed in equal quantities from egg, early blastula (15 hr.), early gastrula (30 hr.) and late gastrula/prism stage (45 hr.) embryos. The methods for sample labeling, array hybridization and washing are identical to those described (*S2*). Arrays were scanned on an Axon 4000B scanner and features were extracted with NimbleScan software (NimbleGen Systems).

Normalization. Data from the individual arrays were normalized on the basis of 5,000 common control probes. First, a reference distribution was created by averaging the sorted signals of the control probes from all 27 arrays. Subsequently, distribution of the control probes from each individual array was mapped to the reference distribution through a fourth-order least-square fit. Finally, signals for all tiling probes were normalized by means of the fitting parameters from the same fourth-order equation.

Example of expressed gene. A typical example (Fig. 1, S1) illustrates the hybridization profiles of an active gene and an adjacent silent gene, superimposed over the predicted gene model structures. The total available range in the units of this normalization is 0-~500. Two factors may account for the data quality: the relatively low biological complexity of the embryo at these stages; and the use of 50-mer rather than shorter probes. The sensitivity of the measurement is shown fora known gene, *sp-foxY* (Fig. S1B). This gene is expressed only in the four small micromeres and later in the archenteron tip. QPCR measurements show that this gene is represented by only 30 mRNA molecules in the whole embryo at 9 h, 125 molecules at 24 h and 400 molecules at 30 and 48 h. Yet every exon is clearly represented in the transcriptional profile.

Analysis of expressed probes. To determine whether the signal on a probe represents real hybridization, a cutoff was chosen based on the random control probes placed on the arrays (1% false positive rate). Genomic probes with signals above the 99% cutoff level (3.7) were considered as expressed. Expressed probes located within 150 bases from the

neighboring ones were combined into contiguous transcriptional units (TU). The transcriptional units matching protein-coding genes were represented by signal from both strands, although the sense strand signals were typically stronger than the antisense. This was an artifact, possibly caused by double-stranded cDNA synthesis during labeling. However, there were 51,081 asymmetric TUs containing two or more array probes, which did not overlap any gene from the predicted set (OGS) (*S1*).

Expression of OGS genes. Determination of OGS genes expressed in embryo was performed with stringent criteria that reduced the number of false positives. All probes located within the OGS genes were considered and probability scores were computed reflecting the likelihood of them being expressed by chance alone. Poisson probability distribution was used. This method identified 12,000-13,000 OGS genes as expressed. The list of expressed genes was further pruned to remove any overlap. All expressed genes were compared with each other using BLAST alignment program (blastp, cutoff 1e-50). Subsequently, the list of matches was filtered to keep only those, for which the length of overlap was more than 90% of at least one of the pair and the ratio of BLAST score and length of overlap was greater than 3.5. The second criterion was used to treat longer and shorter proteins equally. In total, ~1,400 overlaps were determined among the 12,000 expressed genes.

Data availability. All probe sequences and corresponding hybridization data are available from the NCBI GEO database in MIAME-compliant format (accession code GSE6031).

References

S1. Sea urchin Genome Sequencing Consortium, *Science* **314**, 941 (2006).
S2. V. Stolc *et al., Proc. Nat. Acad. Sci. USA* **102**, 4453 (2005).
S3. S. Singh-gasson *et al., Nat. Biotechnol.* **17**, 974 (1999).
S4. E. F. Nuwaysir *et al., Genome Res.* **12**, 1749 (2002).
S5. J. P. Rast, R. A. Cameron, A. J. Poustka, E. H. Davidson, *Dev. Biol.* **246**, 191 (2002).
S6. Q. Tu, C. T. Brown, E. H. Davidson, P. Oliveri, *Dev. Biol.* 10.1016/j.ydbio.2006.09.031 (2006).

Fig. S1. Visualization of transcription profiles in protein coding genes. The protein coding regions of the genes are indicated by the bars, and the orientation of the genes by the DNA strands (W, C) on which they are portrayed. Hybridization of each chip in the array is shown in arbitrary units (ordinate). (**A**) Transcription profile of *Sp-gatac* gene (*S*5). Note the adventitious activity in the single probe in the intron at position about 25,000. The gene is transcribed from right to left and the 3' UTR is visible by comparison to the gene model below. (**B**) *Sp-foxy* gene (*S*6), transcribed from left to right. The wide peak at 73,000-74,000 is the 3' UTR. There are 6 protein coding exons.

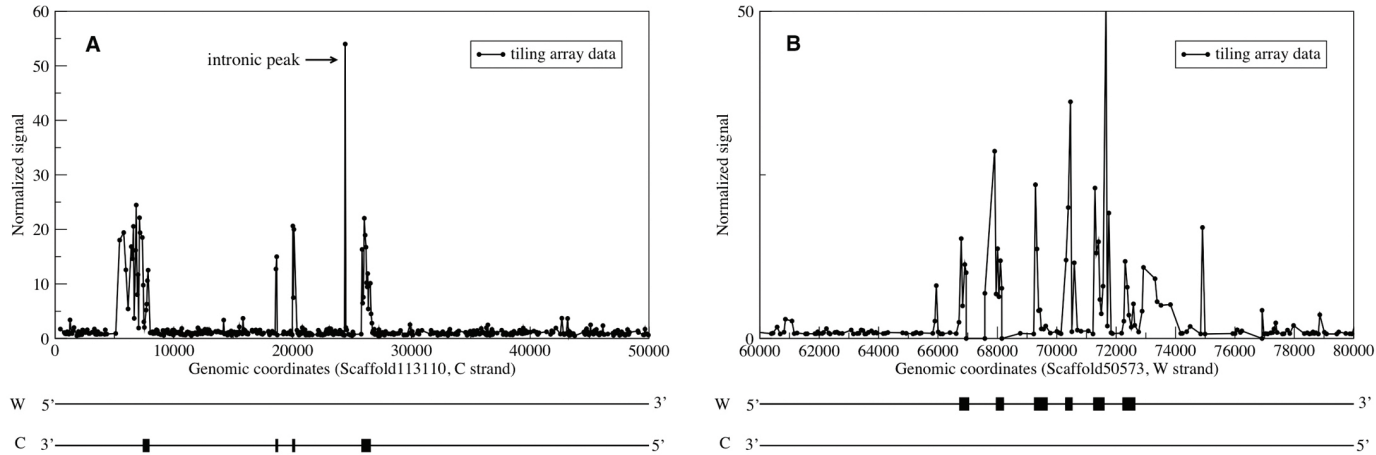
A
intronic peak
tiling array data
Normalized signal
Genomic coordinates (Scaffold113110, C strand)
W 5'
3'
C 3'
5'
B
tiling array data
Normalized signal
Genomic coordinates (Scaffold50573, W strand)
W 5'
3'
C 3'
5'

Table S1. Experimental design and coverage of the genome. In the upper half of the table, numbers for arrays, probe count, coverage of the scaffolds and expressed genes are shown. Lower half presents information for the scaffolds with or without OGS genes. CI, confidence interval

Category	***Number or Size determined***
Total number of arrays	27
Number of tiling probes (unique)	10,187,868 (10,133,868)
Number of scaffolds covered by tiling probes	144,056 (77%)
Sizes of tiling scaffolds (nt)	Mean 7,326, Median 950
Sizes of all scaffolds (nt)	Mean 5,831, median 909
Cutoff based on random probes (95%, 99% CI)	1.56, 3.74
Expressed probes (95%CI)	25% of all probes on the arrays
Expressed OGS genes	~12,500
Non-OGS proteins	~1,000
Asymmetric transcripts	~51,000

	OGS scaffolds*	***Non-OGS scaffolds***
Total length	765,538,835	330,533,581
Number of scaffolds	16,765	171,178
Mean length	45,662	1,931
Median length	21,544	876
Covered by tiling probes	16,360 (98% of GLEAN scaffolds)	127,696 (75% of non-GLEAN scaffolds)
Count of probes	7,632,773	2,555,095
Expressed probes	1,998,385	516,154

*Scaffolds containing OGS genes.

Table S2. Previously characterized genes. The following 28 well-studied embryonic genes were used as a reference in this study. They were manually inspected to confirm the quality of the tiling array data. Their tiling array signals can be visualized at http://www.systemix.org/sea-urchin.

alx1	apobec	bf	bra	capk	Cyclo -phillin	delta
dopt	endo16	eve	foxa	foxb	foxc	gatac
gatae	gcm	gelsolin	kakapo	krox1	lim	neuralized
nk1	nk2-1	not	orct	otx	soxb2	wnt8